Imperial College London

Department of Earth Science and Engineering

MSc in Applied Computational Science and Engineering

Independent Research Project
Project Plan

Prediction of wildfire duration and final burned area with image-based Machine learning

by

Hansong Xiao

Email: hansong.xiao21@imperial.ac.uk
GitHub username: acse-hx221
Repository: https://github.com/ese-msc-2021/irp-hx221

Supervisors:

Dr. Sibo Cheng
Dr. Rossella Arcucci

June 2022

1 Introduction

As humans evolved, fire seems to have become more controllable. But wildfires are not included, they are a natural phenomenon that is difficult to predict and occur with high frequency. More seriously, wildfire is a natural hazard (). Wildfires have a great impact on the world's ecosystems (Bond & Keeley 2005). 4 million km^2 to 6 million km^2 (depends on different estimate method) of land are affected by wildfires every year(Youssouf et al. 2014), which is almost half the area of Europe.

Wildfire has huge negative impact on earth ecosystem, risking world natural system and health, hydrogeomorphic effects from wildfire attack quality of daily water for both human society and ecosystem(Robinne et al. 2018). Wildfire also threatens the world economically, house price of northwest Montana, a potential resort area in USA, was depreciated by unpredictable wildfire in a long-term(Stetler et al. 2010). It also results in serious injury or death for animals(Madigan et al. 2011). To prevent these multiple hazards, researchers have begun to engage in developing with wildfire prediction. Cellular Automata (Chopard & Droz 1998) and CFD (Computational Fluid Dynamic) (Anderson & Wendt 1995) were widely used. However, these models were too computationally intensive, time-consuming, and resource-intensive to solve the wildfire prediction problem well. [Why taking so long] With the development of artificial intelligence, machine learning becomes a new research direction(Fradkov 2020). Satellite imagery was used to help analyze geomorphic features to build models for better prediction.

Most of the ML models were not consistent enough to be widely used. This paper aims to focus on a way to build a fast-decision and general model of wildfire prediction. It will begin with exploring the effect of different features on wildfire prediction by using an image from several databases via the online cloud platform Google Earth Engine (GEE). Currently, four global scale databases were involved in this experiment, providing wildfire information (location, start and end date, fire size, and duration of each fire) (ANDELA et al. 2019), density map of above-ground living biomass storage (Spawn et al. 2020), land cover map of multiple surface characteristics (e.g. grass coverage, tree coverage, snow converge) (Buchhorn et al. 2020), and climate data corresponding to specific location range from 1979 to 2020 (e.g. precipitation, wind speed) (Hersbach et al. 2018). The research for simple regression model was on a local scale, focusing on California, USA because of its representation and attention. The database was split into two parts, from 2003 to 2013 as a train set, and 2013 t0 2015 as a test set. Target area is a square shape distract with the center point of each fire in California, length of the square range from 1km to 30km, specifying the wildfire location from 2003 to 2016. In the first part, traditional Machine Learning models were adopted to illustrate feature importance. In the following part, image-based prediction models will be used, and discussing possible CNN model. The main motivation of this study is to deliver an fast-decision wildfire prediction model and allocate resources rationally.

2 Literature Review

2.1 Traditional Regression Model

2.1.1 Linear Regression

Linear Regression (Montgomery et al. 2021) is a basic data analysing model, which were mostly used in the first place. It could reveal linear, quadratic, and even polynomial relationship between data. Each data contributes to simulate a best fitting curve with smallest cumulative error. Linear Regression has a wide range of application, no matter the number of variables.

2.1.2 K-nearest neighbors algorithm model

K-nearest neighbors (kNN) is a supervised learning algorithm, working on classifier and regression (Altman 1992). kNN assign each data into a class by calculating the Euclidean distance between k nearest point. In regression, the prediction value is mean value of k nearest point.

2.2 Ensemble Learning

2.2.1 Decision Tree

Decision Tree (DT) (Breiman et al. 2017) is a supervised ML algorithm followed by a if-then-else rule, which applied on both classification and regression problem (Breiman et al. 2017). Each node represents an object and each branch indicates one possible output. Hence, a large number of decision node form a decision tree.

2.2.2 Random Forest

Random forest (RF) (Breiman 2001) is a extent of bagging algorithm (Breiman 1996). Bagging is short for Boostrap Aggregation, which is an ensemble method. Bagging combines multiple predictor, based on varies machine learning algorithm, to build an advanced prediction model (Breiman 1996). Bagging aims to train several original DTs on randomly selected subset, and sampling with replacement. For RF, it has one more random step in feature selection at each node. RF sampling both feature and training set, reducing correlation between different decision trees, thus RF normally has better performance than simple decision tree.

2.2.3 Extreme Gradient Boosting)

Extreme Gradient Boosting (Chen & Guestrin 2016) is another ensemble learning algorithm based on Gradient Boost Decision Tree (GBDT) (Ke et al. 2017). XGBoost is more effective and more capable of preventing over-fitting. XGBoost gives regularized term for loss function and fills None value.

3 Problem Description and Objectives

3.1 Object

The project aims to deliver a fast-decision, general model for Wildfire prediction. In the first stage, earth data was extracted from Google Earth Engine, and modify data to a corresponding different target area. It begins with 30km as length, then testing ranges from 1km to 30km by cropping the image of the same proportion. In the next step, different features, labels, and target areas will be tested by four regression methods. The average value will be used for each feature since simple regression cannot deal with the image. These models were unable to give a good prediction by the given mean square error value, which is shown below. Meanwhile, models show inconsistency for different labels (fire size, duration, burned area). The second stage is focused on image-processing by CNN model. Features and target area are based on the results of stage one for an initial test.

3.2 Parameter Tuning

The model is tuned based on some specific parameters, for example, number of estimator was used on random forest. This figure explains mean square error of random forest model converged when number of estimator exceeding 100. More parameter tuning are based on grid search (Bergstra & Bengio 2012).

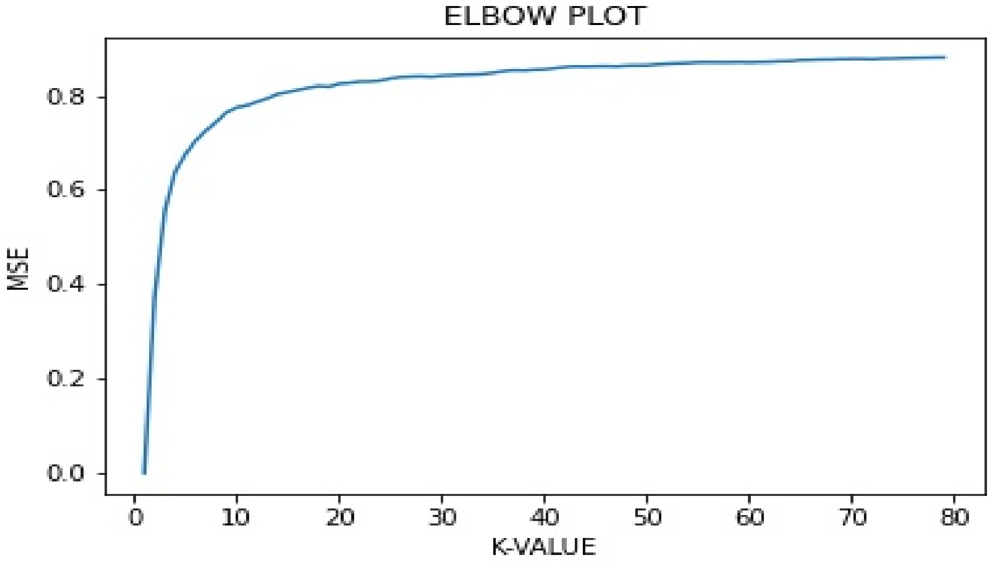

Figure 1: MSE respect to number of estimator

3.3 Regression Model Result

Figure 2. Shows the result of four regression model on the both train and test set on 30km distance. The random forest model shows best performance on train set, opposing on the test set. Other three model preform no big difference.

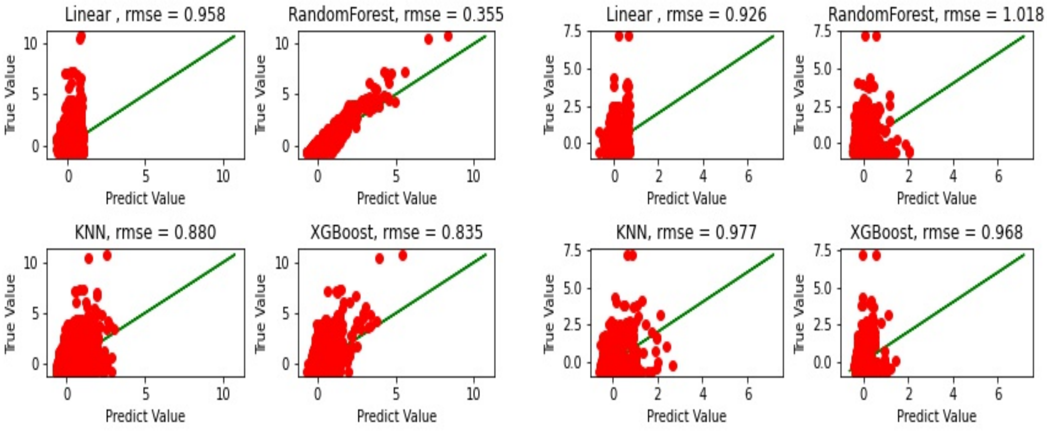

Figure 2: Regression Result of 30 km distance with train set on left, test set on the right

Figure 3. indicates how each model varies with respect to different side length on train and test set.

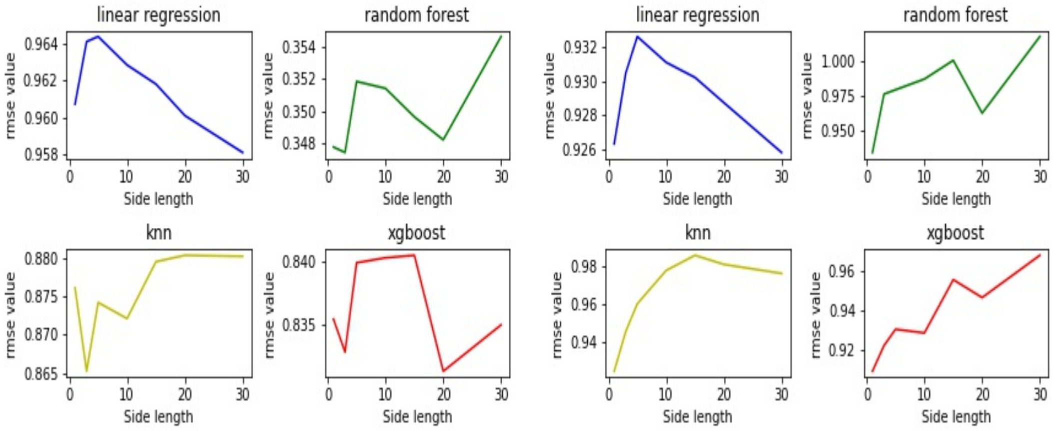

Figure 3: Result of different model ranges from 1km to 30km, with train set on left test set on the right

Figure 4. Feature importance plot on train and test set

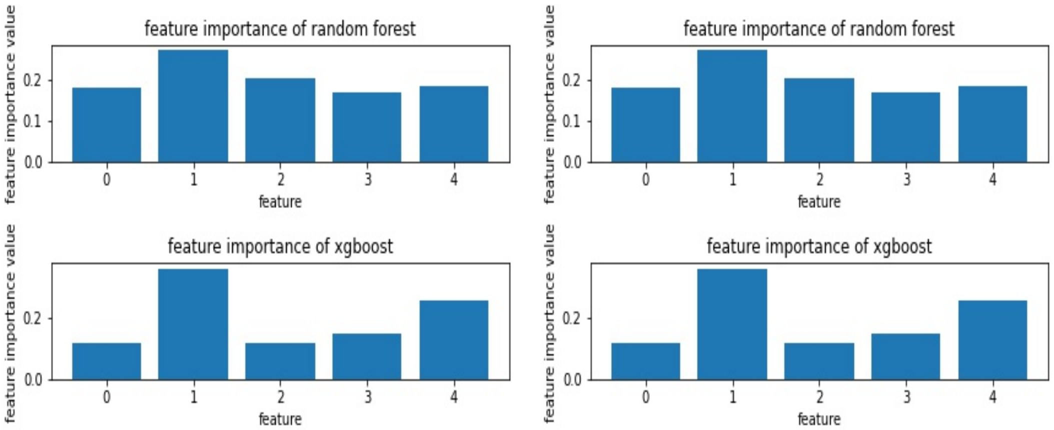

Figure 4: Feature importance with train set on left, test set on the right

4 Progress to Date and Future Plan

Dates	Tasks
6 Jun - 12 Jul	1. Extraction Data from GEE and Global fire atlas dataset 2. Data Pre-processing, getting corresponding coordinates of fires, located CA fires
13 Jul - 20 Aug	1. Implement the simple regression model with different target area and different feature 2. Analysing feature importance, sensitivity analysis, plotting regression result
21 Aug - 28 Aug	1. Writing Project plan, packaging existing model for future use.
1 Jul - 22 Jul	1. Build a CNN Prediction model, train, and test with different hyper-parameters. 2. Compare with the previous regression model.
23 Jul - 13 Aug	1. Implement the final model 2. Analysis results from final model and discuss with accuracy and feasibility
14 Aug - 20 Aug	Testing code, Packaging all model, improving code readability.
21 Aug - 27 Aug	Start writing on report, finishing Github workflow, and final submission.

References

Altman, N. S. (1992), 'An introduction to kernel and nearest-neighbor nonparametric regression', *The American Statistician* **46**(3), 175–185.

ANDELA, N., MORTON, D., GIGLIO, L. & RANDERSON, J. (2019), 'Global fire atlas with characteristics of individual fires, 2003-2016'.

Anderson, J. D. & Wendt, J. (1995), *Computational fluid dynamics*, Vol. 206, Springer.

Bergstra, J. & Bengio, Y. (2012), 'Random search for hyper-parameter optimization.', *Journal of machine learning research* **13**(2).

Bond, W. J. & Keeley, J. E. (2005), 'Fire as a global 'herbivore': the ecology and evolution of flammable ecosystems', *Trends in Ecology Evolution* **20**(7), 387–394.

Breiman, L. (1996), 'Bagging predictors', *Machine learning* **24**(2), 123–140.

Breiman, L. (2001), 'Statistical modeling: The two cultures (with comments and a rejoinder by the author)', *Statistical science* **16**(3), 199–231.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (2017), *Classification and regression trees*, Routledge.

Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L. & Smets, B. (2020), 'Copernicus global land cover layers—collection 2', *Remote Sensing* **12**(6).
URL: *https://www.mdpi.com/2072-4292/12/6/1044*

Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining', pp. 785–794.

Chopard, B. & Droz, M. (1998), 'Cellular automata', *Modelling of Physical* .

Fradkov, A. L. (2020), 'Early history of machine learning', *IFAC-PapersOnLine* **53**(2), 1385–1390.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I. et al. (2018), 'Era5 hourly data on single levels from 1979 to present', *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)* **10**.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T.-Y. (2017), 'Lightgbm: A highly efficient gradient boosting decision tree', *Advances in neural information processing systems* **30**.

Madigan, J., Rowe, J., Angelos, J., Herthel, W., Matz, D., Dinucci, M. & Fletcher, V. (2011), '(a323) wildfire associated burn injury of 1400 sheep in northern california: A coordinated mass casualty veterinary response', *Prehospital and Disaster Medicine* **26**(S1), s90–s91.

Montgomery, D. C., Peck, E. A. & Vining, G. G. (2021), *Introduction to linear regression analysis*, John Wiley & Sons.

Robinne, F.-N., Bladon, K. D., Miller, C., Parisien, M.-A., Mathieu, J. & Flannigan, M. D. (2018), 'A spatial evaluation of global wildfire-water risks to human and natural systems', *Science of The Total Environment* **610-611**, 1193–1206.

Spawn, S. A., Sullivan, C. C., Lark, T. J. & Gibbs, H. K. (2020), 'Harmonized global maps of above and belowground biomass carbon density in the year 2010', *Scientific Data* **7**(1), 112.

Stetler, K. M., Venn, T. J. & Calkin, D. E. (2010), 'The effects of wildfire and environmental amenities on property values in northwest montana, usa', *Ecological Economics* **69**(11), 2233–2243. Special Section - Payments for Ecosystem Services: From Local to Global.

Youssouf, H., Liousse, C., Roblou, L., Assamoi, E., Salonen, R., Maesano, C., Banerjee, S. & Annesi-Maesano, I. (2014), ‘Quantifying wildfires exposure for investigating health-related effects’, *Atmospheric Environment* **97**, 239–251.